Clarity Prediction Challenge 1 Entry: Non-intrusive Speech Intelligibility Metric Prediction - Technical Report

George Close, Samuel Hollands, Stefan Goetze, Thomas Hain

UKRI CDT for Speech and Language Technologies and their Applications, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

{glclose1,shollands1,s.goetze,thain}@sheffield.ac.uk

Abstract

This paper describes an entry to the 1st Clarity Prediction Challenge [1]. Non-intrusive predictors for an intrusive speech intelligibility metric are trained, then fine tuned on the ground truth correctness values in the challenge training data.Results are reported on a number of speech intelligibility metrics, an explanation for the selection of which models have been submit to the challenge is provided.

1. Motivation

In the United Kingdom (UK) 1 in 5, or just over 12 milion people, experience hearing loss of greater than 25 decibels hearing level (dBHL) [2]. By 2035 this will rise to 14.2 million [2] and with age correlating with an individual's likelihood for developing hearing impairment. This statistic is going to inflate dramatically. By 2050 we will have observed a near doubling of the global population aged older than 60 going from just 12% in 2015, to making up 22% of the world's population by 2050 [3], a reality that has large consequences for all medical conditions which increase in likelihood with age.

Inspired by recent works [4, 5] which use a neural network to mimic the performance of an intrusive metric for speech quality and intelligibility, here a similar network structure is used to predict the metric score that will be assigned to the input audio in a 'non-intrusive' way. Note that the technique here differs in that the network is only provided access to the degraded signal, rather than the degraded/clean pair.

We chose to use a metric prediction objective over simply using the ground truth 'correctness' information in the training data as this was found to be distributed in a way that was difficult for our non-intrusive models to find any discernible patterns in. Our intuition is that if these metrics have been found to correlate with human perception of intelligibility, then non-intrusive predictors of said metrics should also. Additionally we report the performance of each of our non-intrusive metric predictors after being 'fine-tuned' on the ground truth intelligibility.

1.1. Speech Intelligibility Metrics

We investigate the non-intrusive prediction of 3 intrusive speech intelligibility metrics, listed here in increasing levels of complexity of computation.

Short-Time Objective Intelligibility (STOI) [6] is a commonly used metric for the assessment of speech intelligibility. It works by computing an average of the correlation between one-third-octave filter-bank representations of the clean and degraded speech signals. It is defined per channel. It has been found to correlate well with human intelligibility in normal hearing individuals [7, 8, 9].

Modified Binaural Short-Time Objective Intelligibility (MBSTOI) [10] is a variant of STOI which takes in binaural (stereo) degraded and reference signals. The score is computed similarly to STOI, except that it includes an internal simulation of the 'better ear effect' wherein the channel with the highest correlation for that block of processing is used to compute the final score.

The Hearing-Aid Speech Perception Index (HASPI) [11] metric is designed specifically to assess intelligibility in people with hearing loss. In addition to a degraded and a reference signal it also takes an audiogram representation of the hearing loss in a given ear, and incorporates a hearing loss simulation as part of the computation of the score. It additionally incorporates an ensemble of neural networks fitted to real human intelligibility as part of the score calculation.

2. Experiments

2.1. Tools and Software

We implement our experiments via modifications to the challenge baseline system, replacing the simple fitting model with the neural models described below using PyTorch [12]. We also use some features of the SpeechBrain [13] package for audio loading and dataloader creation. For the computation of the STOI scores we use a Python implementation, for MBSTOI we use the Python implementation provided in the baseline and for HASPI we use the MATLAB implementation provided in the challenge documentation. All of the models are relatively low cost, and can be run on a CPU in a reasonable amount of time.

2.2. Feature Extraction

Features are calculated from the time domain hearing aid processed signals either the output of the hearing aid x or $\hat{x}$ the output of the hearing aid processed by the hearing loss simulation [14] used in the baseline system, depending on the metric. First a spectral magnitude $X_{k,\ell}$ for frequency k and frame ℓ is calculated of the time domain audio signal X, followed by a transformation to the feature space by adding 1 to and taking the logarithm of each element to give the feature representation $\mathbf{X}_f$. The channel index c is denoted using $\mathbf{X}_f^c$. Note that in the following $\mathbf{X}_f^c$ denotes the feature representation of the hearing aid output while $\hat{\mathbf{X}}_f^c$ is the feature representation of the hearing aid output x with the baseline hearing loss applied $\hat{x}$.

2.3. Model Structure for Non-Intrusive Prediction

For each of the 3 metrics we investigate, we adapt the same basic model structure for the specific requirements of the metric. The basic structure is based on that of the discriminator network in [5] - 4 2D convolutional layers with 15 filters of a kernel size of (5, 5). To account for the variable length of input data,

a global 2D average pooling layer is placed immediately after the input, fixing the feature representation at 15 dimensions. After the convolutional layers, a mean is taken over the 2nd and 3rd dimensions, and this representation is fed into 3 sequential linear layers, with 50, 10, and 1 output neuron(s) respectively. The first 2 of these layers have a LeakyReLU activation while the final layer has no activation.

For STOI, we predict the score for each channel of the HA output audio separately, with the input to the prediction network being the feature space representation of the given channel $\hat{\mathbf{X}}_f^c$ where c is a channel index. As such, the input dimension to the average pooling and first 2D convolutional layer is 1.

For MBSTOI, we predict the score for the HA output stereo audio together, with the input to the network being the feature space representations of both channels $\{\hat{\mathbf{X}}_f^l, \hat{\mathbf{X}}_f^r\}$. The input dimension of the average pooling layer and the initial convolutional layer is 2 to account for these stacked channel representations.

Finally for HASPI which like STOI is defined per audio channel, we use the $\mathbf{X}_f^c$ representation of the audio, but also use $\mathbf{a}^c$ the audiogram representation of the listener's hearing loss for channel c as input. This 6 element representation is passed through a linear layer with 10 output neurons then another with 50; this representation is then concatenated along the feature dimension with the representation of the audio of the same size. This 100 element representation is then fed through a further 3 linear layers with 50, 10, and 1 output node(s) respectively, all but the last layer having a LeakyReLU activation.
Additionally, we train a model with the same structure as that for the HASPI prediction described above, and train it to predict the ground truth Correctness scores in the training data.

2.4. Experiment Setup

We pre-compute the STOI, MBSTOI, and HASPI scores for the entire train set. We then train 4 models, as described above, to reproduce the score, given only the hearing aid output with hearing loss simulation $\hat{\mathbf{X}}_f^c$ for STOI and MBSTOI, and in the case of HASPI and the model directly predicting the ground truth Correctness, the hearing aid output $\hat{\mathbf{X}}_f^c$ and the audiogram information $\mathbf{a}^c$.
We use a Short Time Fourier Transform (STFT) with a window length of 20ms, a hop length of 10ms and an FFT size of 1024. The hearing aid outputs x have a sampling rate of 32kHz, while the hearing aid outputs with the baseline hearing loss simulation applied $\hat{x}$ have a sampling rate of 44kHz.
Following on from the baseline system we train with a 5 fold validation technique, partitioning the folds on the scene ID. We use the Adam [15] Optimiser with a learning rate of 0.001 for all models.
All models are trained with a batch size of 1 with the exception of the model that directly predicts correctness which uses a batch size of 20. We additionally 'fine tune' the metric objective models using the ground truth 'correctness' (intelligibility) scores; in the case of the the metrics that are defined per channel (STOI and HASPI) we use the channel that returned the highest predicted score between the 2, as a simplified simulation of the 'better ear effect'. This finetuning process consists of exposing the model to the entire training set in the same way as in the pre-training, but having it's outputs compared to the ground truth rather than the metric. We use this same technique to evaluate the performance of these models. Research into optometry demonstrates that grammatically intact sentences provide a lower accuracy for diagnosis than a bag of words approach [16] as language cognition interferes with the results. In the context of hearing loss it seems likely language cognition will also have an impact and therefore the corpus used containing grammatically intact sentences may have hindered the performance of the metrics evaluated.

Table 1: *Performance on the Clarity Prediction Challenge Train Set*

Model Objective	**Correctness Error**	r	ρ
STOI	35.63	0.3	0.21
STOI (fine)	34.55	0.32	0.25
MBSTOI	39.30	0.26	0.18
MBSTOI (fine)	34.72	0.32	0.23
HASPI	38.80	0.23	0.22
HASPI (fine)	**31.55**	**0.53**	**0.46**
Correctness	33.44	0.45	0.42
	Prediction Error	r	ρ
STOI	**13.88**	0.43	0.3
STOI (fine)	16.44	0.43	0.3
MBSTOI	15.50	0.44	0.33
MBSTOI (fine)	21.81	0.47	0.32
HASPI	25.10	**0.59**	**0.59**
HASPI (fine)	37.09	0.29	0.29

2.5. Results

Table 1 shows the results of our experiments over the entire training set for the challenge. The upper half shows the Root Mean Square Error (RMSE) of the model outputs versus the ground truth 'correctness' values. The lower half shows the RMSE versus the true values for the target metric of the model i.e the prediction error. r and ρ are the Spearman and Pearson Correlations respectively.
In terms of prediction error, the model that is best able to non-intrusively predict it's target metric is that which is the STOI prediction model, while the worst is the HASPI model. This is likely because the calculation of the STOI is considerably simpler than that for HASPI. As is to be expected, the fine tuning to the ground truth correctness has the effect of increasing the prediction error while decreasing the correctness error for all models.
The best model in terms of prediction of the ground truth correctness was the fine tuned HASPI predictor. The slight performance improvement versus the model that was only trained to predict the correctness shows that the HASPI objective pre-training did improve performance.

3. Conclusion

Of the models we have trained, we selected the Fine-tuned HASPI objective (`E006`) and the Correctness (`E034`) objective models as submissions to the challenge, as they are the best performing over the training set.

4. References

[1] S.Graetzer, J.Barker, T.J.Cox, M.Akeroyd, J.F.Culling, G.Naylor, E.Porter, and R. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing."

[2] N. Park, "Population estimates for the UK, England and Wales, Scotland and Northern Ireland, provisional: mid-2019," *Hampshire: Office for National Statistics*, 2020.

[3] W. H. Organization, *World report on ageing and health*. World Health Organization, 2015.

[4] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," 2018.

[5] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," 2021.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[7] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[8] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, Sep. 2014.

[9] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors," in *2020 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, October 2020, pp. 6493–6497.

[10] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639317302947

[11] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Communication*, vol. 65, pp. 75–93, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000545

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[13] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[14] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997. [Online]. Available: https://doi.org/10.1121/1.419733

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[16] M. MacKeben, U. K. Nair, L. L. Walker, and D. C. Fletcher, "Random word recognition chart helps scotoma assessment in low vision," *Optometry and Vision Science*, vol. 92, no. 4, p. 421, 2015.